# Supervised Learning Approach for Predicting the Quality of Cotton Using WEKA

M. Selvanayaki, M.S. Vijaya, K.S. Jamuna, and S. Karpagavalli

PSGR Krishnammal College for Women
Coimbatore – 641 004, India
{msvijaya,ksjamuna,karpagam}@grgsact.com

**Abstract.** Cotton is the world's most important natural fibre used in Textile manufacturing. Cotton fiber is processed into yarn and fabric. Yarn strength depends extremely on the quality of cotton. The physical characteristics such as fiber length, length distribution, trash value, color grade, strength, shape, tenacity, density, moisture absorption, dimensional stability, resistance, thermal reaction, count, etc., contributes to the quality of cotton. Hence determining the quality of cotton accurately is an essential task to make better raw material choices in textile industry which in turn will support better buying and selling decisions. In this work, cotton quality prediction is modeled as classification task and implemented using supervised learning algorithms namely Multilayer Perceptron, Naive Bayes, J48 Decision tree, k-nearest neighbor in WEKA environment on the cotton quality assessment dataset. The classification models have been trained using the data collected from a spinning mill. The prediction accuracy of the classifiers is evaluated using 10-fold cross validation and the results are compared. It is observed that the model based on decision tree classifier produces high predictive accuracy compared to other models.

**Keywords:** Machine learning Techniques, Multilayer Perceptron, Naïve Bayes, J48, k-Nearest Neighbor.

## 1 Introduction

Cotton is the most commonly used textile fiber in the world. Its current market share is 56 percent for all fibers used for apparel and home furnishings. Cotton fiber is spun into yarn before being woven or knitted into fabric. The quality of the cotton fibre is determined by three factors, namely, the color of cotton, purity (the absence of foreign matter) and quality of the ginning process, and the length of fibers.

The color of cotton fibres is primarily determined by conditions of temperature and/or humidity, cotton lint exposure to sunlight, and cotton varieties. Fibre length is defined as the average length of the longer one-half of the fibres (upper half mean length) [1]. Length uniformity or uniformity ratio is determined as a ratio between the mean length and the upper half mean length of the fibers and is expressed as a percentage. Fiber strength is measured in grams per denier. Micronaire measurements reflect fiber fineness and maturity. A constant mass 2.34 grams of cotton fibers is compressed into a space of known volume and air permeability measurements of this

V.V Das et al. (Eds.): BAIP 2010, CCIS 70, pp. 382–384, 2010.

compressed sample are taken. These, when converted to appropriate number, denote Micronaire values [2]. A trash measurement describes the amount of non-lint materials in the fiber. Trash content is highly correlated to leaf grade of the sample. There are seven leaf grades ranging from #1 to #7 and a lower grade #8.

United States Department of Agriculture (USDA) classification specifically identifies the characteristics of fiber length, length uniformity, strength, Micronaire, color, preparation, leaf and extraneous matter. This work concentrates on the physical properties of cotton for prediction by employing machine learning techniques.

Machine learning provides methods techniques and tools, which help to learn automatically, and to make accurate predictions based on past observations. Machine learning is popularly being used in areas of business like data analysis, financial analysis, stock market forecast and so on. The supervised learning techniques namely k-nearest neighbor, Multilayer Perceptron, Naïve Bayes Classifier, Decision tree induction have been used to learn and build the classification models.

## 2 Experimental Setup

The data set with 12 different features was collected from a private spinning mill. The training was carried out with 14325 instances. The predominant features that decide the quality of cotton include span length (mm), uniformity ratio%, strength (g/tex), micronarie, lint, trash, invisible loss, maturity coefficient. As the quality of the cotton has to be predicted, Quality_Indicator is selected as the class label. The instances in the dataset pertaining to the three quality measures of cotton are labeled as Low (L), Medium (M) and High (H). Using WEKA [3], a software environment for machine learning, independent models have been trained for predicting the cotton quality. The performance of the classifiers is evaluated and the results are analyzed. The 10-fold cross validation has been applied to test the performance of the four models on cotton data.

## 3 Results and Discussion

The statistical results of the experiments are summarized in Table 1. The performances of the four models are evaluated based on the three criteria, the prediction accuracy, learning time and error rate.

The prediction accuracy of the four models is shown in Fig 1. J48 algorithm shows the higher accuracy than the other models. The accuracy rate of the Naive bayes model is comparatively very low. The learning time of the four schemes under consideration are shown in Fig. 2. Multilayer perceptron consumes more time to build the model. The Naïve Bayes classifier and KNN learn more rapidly for the given dataset. The error rate for J48 is insignificant thereby indicating higher accuracy than the other models. The Multilayer perceptron model produces higher error rate than the other machine learning algorithms employed in this work.

**Table 1.** Predictive Performance of the Classifiers

| Evaluation Criteria | J48 | NB | MLP | NN |
|---|---|---|---|---|
| Kappa Statistics | 0.9827 | 0.2841 | 0.8524 | 0.9827 |
| Mean Absolute Error | 0.0048 | 0.2487 | 0.0167 | 0.0042 |
| Root Mean Squared Error | 0.0491 | 0.4305 | 0.0094 | 0.0463 |
| Relative Absolute Error | 2.3237% | 121.52% | 8.1462% | 2.0594% |
| Root Relative Squared Error | 15.352% | 134.59% | 31.074% | 14.464% |
| Time to build the model (sec) | 0.3 | 0.17 | 226.58 | 0.8 |
| Correctly classified instance | 14239 | 8871 | 13726 | 14239 |
| Incorrectly classified instance | 76 | 5444 | 589 | 76 |
| Prediction Accuracy | 99.46% | 61.97% | 95.88% | 99.46% |

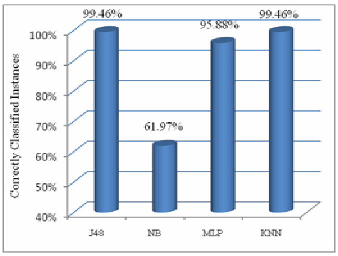


**Fig. 1.** Prediction Accuracy

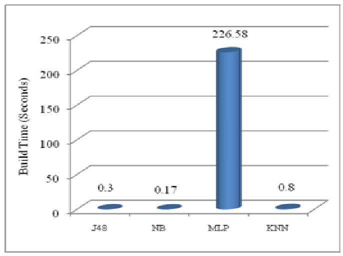


**Fig. 2.** Learning Time

## 4 Conclusion

In this work, cotton quality prediction is modeled as classification task and four supervised machine learning schemes namely K-Nearest Neighbor, Multilayer Perceptron, Naïve Bayes Classifier, Decision tree induction have been applied on the cotton data in order to learn and predict the quality of cotton. The independent trained models are generated and their performance has been evaluated based on their predictive accuracy and ease of learning. The results indicate that the J48 classifier outperforms other classifiers.

## References

1. Lawrence, H.S.: Cotton's Importance in the Textile Industry. In: Symposium, Lima, Peru (1998)
2. Gordon Cook, J.: Handbook of Textile Fibers, Part I. Natural Fibers. Merrow Publishing Co. Ltd. (1968)
3. Witten, I.H., Frank, E.: Data Mining – Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, Amsterdam (2005)